Lignocellulolytic carbohydrate-active enzyme-encoding gene clusters in novel *Trichoderma atroviride* and *Trichoderma harzianum* strains: A BAC-guided transcriptome analysis

Paulo Henrique Campiteli[a, b], Rafaela Rossi Rosolen[a, b], Maria Augusta Horta[a, c], Juliano Sales Mendes[a, d], Carla Cristina da Silva[a, d]; Danilo Sforça[a], and Anete Pereira de Souza[a, e]

h.paulocampiteli@gmail.com, rafaelarosolen@gmail.com, mariaaugusta.horta@gmail.com, juliano16@gmail.com, carla.silva1@ufv.br, danaugus@unicamp.br, anete@unicamp.br

[a]Center for Molecular Biology and Genetic Engineering (CBMEG), University of Campinas (UNICAMP), Cidade Universitária Zeferino Vaz, Campinas, SP, Brazil - 13083-875

[b]Graduate Program in Genetics and Molecular Biology, Institute of Biology, UNICAMP, Campinas, SP, Brazil – 13083-970

[c]State São Paulo University (UNESP) – Estrada Municipal do Campinho, s/n - Pte. Nova, Lorena - SP, 12602-810

[d]State University of Viçosa (UFV) – Av. P H Rolfs, s/n - Campus Universitário, Viçosa - MG, 36570-900

[e]Department of Plant Biology, Institute of Biology, UNICAMP, Cidade Universitária Zeferino Vaz, Rua Monteiro Lobato, Campinas, SP, Brazil - 13083-862

Abstract

Lignocellulosic biomass is a challenging substrate for enzymatic hydrolysis and therefore increases the process cost in biorefineries. In nature, filamentous fungi naturally degrade lignocellulose by using an arsenal of hydrolytic and oxidative enzymes, specifically, carbohydrate-active enzymes (CAZymes). These fungi, including those of the genus *Trichoderma*, produce and secrete an arsenal of CAZymes that act synergistically to efficiently degrade complex and recalcitrant substrates. *Trichoderma atroviride* CBMAI-0020 and *Trichoderma harzianum* CBMAI-0179 are biocontrol strains against important phytopathogens and naturally degrade lignocellulosic substrates. We employed the bacterial artificial chromosome (BAC) methodology together with gene expression data under cellulose degradation conditions to identify novel degradative enzymes, explore their proximal context, and identify CAZyme gene clusters. Using this simple, time-saving, and cost-effective molecular tool, it was possible to identify genomic regions rich in CAZymes. We found important gene clusters related to lignocellulose degradation by inspecting the linear structure and expression data of CAZyme-rich regions. The results offered a genomic perspective of important regions related to carbohydrate metabolism, uncovering novel targets positively regulated in the degradation of lignocellulose and helping to elucidate differences in biomass degradation among *Trichoderma* species.

Keywords: *Trichoderma*, bacterial artificial chromosomes (BACs), structural genomics, carbohydrate metabolic gene clusters, multiomics

Abbreviations

BCA – Biocontrol agents

CAZyme – Carbohydrate-active enzymes

TF – Transcription Factor

TR – Transporter

SMBGC – Secondary Metabolism Biosynthesis-related Gene Cluster

BAC – Bacterial Artificial Chromosome

PDA – Potato Dextrose Agar

Introduction

Trichoderma spp. (ascomycetes) are opportunistic generalists that are known to be efficient plant biomass decomposers (Druzhinina et al., 2018, Kubicek et al., 2019). They are distributed worldwide and colonize soil, living plants, decomposing wood, and other fungi (Srivastava and Shahid, 2014 Zeilinger et al., 2016; Druzhinina et al., 2018; Kubicek et al., 2019). These species have previously been of interest in biotechnology due to their efficient lignocellulolytic activities, especially *Trichoderma reesei* QM6a (and its derivative clones) (Li et al., 2017; Li and Wang, 2021). They are widely used for the industrial production of cellulases and hemicellulases for enzymatic cocktails, which are applied in saccharification (Bischof et al., 2016) and are used to optimize enzymatic hydrolysis processes in biorefining. In addition, the emerging interest in *Trichoderma* is linked to the remarkable biological activities of these species, with outstanding biocontrol and biofertilization capacities. They are capable of controlling rhizospheric microorganisms while displaying beneficial interactions with plants and are considered prominent biocontrol agents (BCAs) that pose low risk to humans (Kubicek et al., 2019; Sood et al., 2020). *Trichoderma atroviride* and *Trichoderma harzianum* are mycoparasites displaying biocontrol activities against various phytopathogens, including *Rhizoctonia* sp., *Botrytis* sp., *Fusarium* sp., and *Penicillium* sp. (Sood et al., 2020; Stracquadanio et al., 2020; Nascimento et al., 2022). Additionally, plant-*Trichoderma* cocultures exhibit biofertilization activities, promoting vegetative traits and immunologic defense systems and attenuating abiotic stress (Kubicek et al., 2011, 2019; Sood et al., 2020; Tyśkiewicz et al., 2022; Guzmán-Guzmán et al., 2023). Members of the genus *Trichoderma* possess a remarkable system of substrate perception and are able to trigger specific reactions for metabolic adaptation and to synthesize and transport proteins related to lignocellulolytic activity, such as carbohydrate-active enzyme

(CAZyme) mixtures, transcription factors (TFs), and transporters (TR). They also produce other complex metabolites for defense, competition, or interaction with other organisms, such as antibiotics or antimycotics (Andlar et al., 2018). CAZymes are enzymes related to cleavage, synthesis, or interaction with carbohydrate polymers, as listed in the Carbohydrate-Active Enzyme database **(**http://www.cazy.org/**)** (Andlar et al., 2018; Drula et al., 2022).

The CAZyme composition and hydrolytic reactions promoted when *T. harzianum* CBMAI 179 (TH0179) and *T. atroviride* CBMAI 0020 (TA0020) were under biomass degradation conditions were previously investigated. Their cellulolytic potential was evaluated by quantifying cellulase and hemicellulase production through filter paper activity tests (Horta et al., 2018). The transcriptomic profiles of both strains were explored by cultivation under cellulose degradation conditions (Almeida et al., 2021). The results indicated TH0179 to be an efficient cellulose degrader, while TA0020 showed a complex transcriptomic profile with a lower lignocellulolytic potential (Horta et al., 2018; Almeida et al., 2021; Rosolen et al., 2022). The heterogeneous responses to cellulose are indications of the high complexity of the genus, in which closely related species carry relevant species-specific traits affecting complex biological processes.

Bacterial artificial chromosomes (BACs) are a classic molecular biology tool used to recover and study target genomic regions (Peterson et al., 2000). The entire genome is distributed in libraries, with a rapid selection platform used for the identification of specific gene targets by PCR. This technique can be applied for gap closure in fragmented assemblies, as an assembly method (as described for the first human genome sequence (Waterston et al., 2002), or as a tool for studying targeted genomic regions in higher eukaryotes with complex genomes rich in polyploid and aneuploid repetitive elements (Visendi et al., 2016; Mancini et al., 2018). In

summary, it is a simple, reliable, low-cost method for studying target genomic regions and is especially powerful for studying novel species without previous genomic data.

In fungi, the BAC approach has been successfully applied to investigate genomic regions with CAZyme cluster genes in *T. harzianum* IOC 3844 (Filho et al., 2020), (Crucello et al., 2015). In *Streptomyces* spp., for example, the BAC approach was used for genome mining and revealed important gene clusters of bioactive natural products and the corresponding biosynthetic pathways (Xu et al., 2016). Another study used BAC libraries to characterize a novel gene conferring sensitivity to the necrotrophic fungal pathogen *Parastagonospora nodorum* (Zhang et al., 2021).

In this study, a rapid selection platform was used. Using the BAC approach, biomass-related genomic regions were investigated in the TH0179 and TA0020 strains. This methodology allowed us to access genomic regions that encode cellulases, hemicellulases, and ligninases, as well as TFs, transporters, and potential targets involved in cellulose degradation in the evaluated strains. In addition, we employed the data generated from a previous RNA-Seq study (Almeida et al., 2021) to analyze the expression of CAZyme genes through a BAC-guided transcriptomic analysis, which allowed the identification of lignocellulolytic CAZyme clusters that are positively regulated for the degradation of crystalline cellulose.

The results offer insights into the genomic organization and clustering of CAZymes in novel *Trichoderma* strains. The gene content of the genomic region, linear organization, and proximal context was integrated with transcriptomic data to identify lignocellulolytic CAZyme clusters and start addressing observed differences in lignocellulolytic activities. Additionally, novel genes without any functional annotation were expressed in the presence of cellulose; therefore, these were identified as species-specific gene targets for further characterization.

Material and methods

Strains and data availability

TH0179 and TA0020 were provided by the Brazilian Collection of Environment and Industry Microorganisms (CBMAI), located in the Chemical, Biological, and Agricultural Pluridisciplinary Research Center (CPQBA) of the University of Campinas (UNICAMP), Brazil. The identity of the *Trichoderma* isolates was authenticated by CBMAI based on phylogenetic studies of their internal transcribed spacer (ITS) region and translational elongation factor 1 (tef1) marker gene.

The CAZyme target sequences were obtained from the *T. harzianum* T6776 (Baroncelli et al., 2015) reference genome for TH0179 and the *T. atroviride* IMI206040 (Kubicek et al., 2011) reference genome for TA0020 and used as references for CAZyme consensus sequences for RNA-Seq read mapping. TH0179 and TA0020 gene expression data were obtained in a previous study (accession number PRJNA336221) (Almeida et al., 2021), in which each fungal species was cultivated with crystalline cellulose and glucose as the only carbon sources in each condition. Details on the culture conditions are provided in Supplementary Material S1.

BAC library construction

The BAC libraries consisted of 5760 individual clones, each carrying a unique long genomic fragment (20 kb to 200 kb), constructed following (Peterson et al., 2000), with minor modifications for filamentous fungal samples. For each species, 10 µl of spore culture was inoculated on potato dextrose agar (PDA) plates (100 µg.ml-1 ampicillin, 34 µg.ml-1

chloramphenicol) at 28°C for four days. Next, a 2 cm x 2 cm portion of the colony was inoculated on sterilized potato dextrose and incubated (28°C, 3 days, 200 rpm). The final liquid culture was filtered using sterile *Miracloth* (Calbiochem, USA), and the mycelia were used for high-molecular-weight genomic DNA (HMW-gDNA) extraction using the Qiagen DNeasy Plant Mini Kit II (Hilden, Germany). The HMW-gDNA was immobilized on low melting agarose plugs for BAC library construction, and large genomic DNA inserts were cloned and inserted into *E. coli* using the pAGI BAC vector (provided by CNRGV). Inserts were transformed using *E. coli* DH10B competent cells by electroporation.

The transformed culture was inoculated on PDA plates, and the successfully inserted fragments were identified visually in white colonies. Each individual colony represented a unique clone and was cultured and plated on 384-well plates. Each library was examined on a rapid selection platform composed of combined rows or column pools, which allowed fast screening of the library through real-time qPCRs of target genes using a previously described method (Crucello et al., 2015; Filho et al., 2020) (**Supplementary material S2**).

Screening of CAZyme genes and BAC-DNA extraction

At the time, no previous genomic data were available for the target strains. Genome sequences of *T. harzianum* T6776 (Baroncelli et al., 2015) were used as the reference genome and for primer construction in Primer3Plus (Untergasser et al., 2012) for the TH0179 library screen, and *T. atroviride* IMI 206040 (Kubicek et al., 2011) was used as a reference for the TA0020 library screen. The primers were used for amplification on the rapid selection platform, (Crucello et al., 2015; Filho et al., 2020) and BACs were selected via melt curve comparison. To avoid nonspecific and false positive amplification, the final selection was performed by 3%

agarose electrophoresis (80 V, 2 h, 1 kB Promega ladder). The selected clones were preincubated in growth medium (15 ml, 200 rpm, 37°C, 6 h-12 h) and then incubated overnight in 80 ml of growth medium (200 rpm, 37°C, 6 h-12 h). After incubation, the culture was centrifuged, and BAC-DNA was extracted using a Macherey-Nagel Plant Mini Kit II (Düren, Germany) following standard procedures. Each positive extracted clone was subjected to quality control as described in **Supplementary material S2**, and **Supplementary material S3** corresponds to the Hind-III BAC digestion agarose 1% electrophoresis and agarose 3% electrophoresis for pool integrity checking.

BAC-DNA assembly and evidence-based gene prediction

The sequencing data were transferred to a local server for polishing and assembly. Raw reads were polished using Quiver, and BACs were *de novo* assembled using Canu v.2.1.1 (Koren et al., 2017).

Canu-assembled genomic regions were aligned to the *T. harzianum* T6776 (Baroncelli et al., 2015) and *T. atroviride* IMI 206040 (Kubicek et al., 2011) reference genomes to filter *Trichoderma* sequences from the respective BAC sequencing bulks. BACs <40 kb were filtered out of the analysis. Then, all-vs.-all alignment was performed for TA0020 and TH0179 BACs individually to select nonoverlapping regions.

For evidence-based gene prediction, we first aligned RNA-Seq reads to BACs using STAR (Dobin et al., 2013). BAM outputs were inputted into StringTie 2.2.1 (Pertea et al., 2015) for transcript assembly guided by the BAC sequences. The final set of transcripts was used as input for maker2 2.31.11 (Holt and Yandell, 2011) gene structure prediction. The predicted

models were used for further functional inference and comparative analysis. The final GFF gene models are available in **Supplementary material S4**.

Functional annotation and cluster identification

Gene models were processed with InterProScan v5.55-88.0 (Quevillon et al., 2005; Jones et al., 2014) and eggnog-mapper v.2.0.1 (Huerta-Cepas et al., 2019). Parallel annotation of the CAZyme modules and CAZyme clusters was conducted with the dbCAN v3 web platform (Zheng et al., 2023), and KOALAblast (Kanehisa et al., 2016) was used to identify metabolism-related genes from http://www.kegg.jp (Kanehisa and Goto, 2000) with a search against *Hypocreceae*. The antiSMASH web server 6.0 (Blin et al., 2021) was used to define secondary metabolite biosynthesis-related gene clusters (SMBGCs) (TA0020 output available in **Supplementary material S9**, TH0179 in **Supplementary material S10**). Functional data and gene structure prediction were integrated using Funannotate (Palmer and Stajich, 2020; Li and Wang, 2021) implemented on the Galaxy platforms (Galaxy Community, 2022). The functional data and differentially expressed gene (DEG) information are available in **Supplementary material S5**, and the putative CAZymes annotated via dbCAN3 are available in **Supplementary material S6**.

Carbohydrate degradation gene expression analysis

Previously published gene expression data (Almeida et al., 2021) obtained under 96 h-cellulose and glucose fermentation conditions were used for expression inference and differential expression analysis. The selected genomic regions from TH0179 and TA0020 were used to perform BAC-guided transcriptome assembly. The FastQC v0.11.5 (Trivedi et al., 2014) and

Trimmomatic 0.39 (Bolger et al., 2014) parameters were as follows: min length 36 bp, SLIDINGWINDOW:4:10:10, used to access and filter reads by quality. Filtered reads were aligned to BAC sequences on STAR 2.7.10b aligner (Dobin et al., 2013) using the final merged annotation as a reference, and the BAM output was used to perform referenced-based StringTie 2.2.1 transcript assembly (Pertea et al., 2015). In addition, the TPM value for each gene was calculated automatically based on StringTie quantification, and the TPM values are available in **Supplementary material S5**. The EdgeR package (Robinson et al., 2010) was used to compare expression under the studied conditions and identify genes with significant differential expression by a pairwise comparison of annotated gene models, with glucose considered the control.

Comparative analysis

Reference alignment

For the structural comparison, BACs were aligned to the *T. atroviride* IMI 206040 (Kubicek et al., 2011) and *T. harzianum* CBS226.95 (Baroncelli et al., 2015) draft genomes and the *T. reesei* QM6a (Li et al., 2017) reference genome using MUMmer (Marçais et al., 2018). The MUMmer dotplot parameters employed to visualize the alignment were as follows: all-vs.-all, minimum match 50 bp, and min alignment length 100 bp.

Orthology

Newly annotated gene models were compared to 24 *Trichoderma* species plus *Saccharomyces cerevisiae* S288C (Goffeau et al., 1996), *Neurospora crassa* OR74A (Galagan et

al., 2003), *Pleurotus ostreatus* PC15 (Riley et al., 2014), and *Phanerochaete chrysosporium* RP-78 (Ohm et al., 2014) as external groups, all extracted from the mycocosm (Grigoriev et al., 2014) (the list of species is available in **Supplementary material S7**), using OrthoFinder v2.5.2 (Emms and Kelly, 2019). OrthoFinder comparative genetics and per-species output are available in **Supplementary material S8**.

Theory

This BAC-guided transcriptomic analysis is based on the natural lignocellulolytic activities observed in *Trichoderma* species, in which individual species can carry species-specific traits affecting relevant activities, such as biomass degradation (Kubicek et al. 2019). *Trichoderma* genomic-based studies have already shown a clustering pattern of CAZymes, which increases the relevance of the BAC approach for studying novel strain degradative enzymes from a genomic point of view (Crucello et al. 2015, Filho et al. 2020). Finally, the BAC approach offers high-quality genomic data with high accuracy to explore interest-specific regions for comparing gene arrangements among different species and revealing the evolution of their genome structure. Therefore, based on such characteristics, the BAC approach was used to mine genomic regions using CAZymes from closely related species and explore their vicinities to recover clustered CAZymes, TFs, TRs, and other genes, such as SMBG, which can impact *Trichoderma*'s degradative activities.

Results

BAC screen

The final library consisted of 5760 individual clones, each carrying a unique long genomic fragment (20 kb to 200 kb) and was examined using a rapid selection platform with plates containing pools for rapid screening of the whole library. For genomic region screening, (I) 25 and (II) 31 CAZyme primers selected (I) 38 and (II) 63 BAC clones for TH0179 and TA0020, respectively (the results are available in **Supplementary material S2**). Only BACs larger than 40 kb were considered for further analysis. Then, all-vs.-all alignment was performed for each strain, and we filtered a final set of 24 nonredundant genomic regions for TH0179 and 44 for TA0020, ranging in size from 41 kb to 138 kb and 53 kb to 164 kb, respectively.

Assembly and genomic characterization

The CAZyme target transfer from *T. harzianum* T6776 for TH0179 was less successful than that from *T. atroviride* IMI 206040 for TA0020, and only 24 unique regions were selected after assembly and filtering, accessing 1.9 Mb, ~5% of the *T. harzianum* CBS226.95 genome. TA0020, on the other hand, recovered 44 BACs, reaching 4.7 Mb, 15% of the *T. atroviride* IMI 206040 genome (Table 1).

Multiple tools were used to access functional data for the annotated models. Ninety-three percent of the TA0020 annotated gene model and 91% of the TH0179 annotated gene model had at least one functional annotation. CAZyme modules represented 9.3% of the TH0179 annotated gene model and 11% of the TA0020 annotated gene model. AntiSMASH recovered SMBGCs physically linked to important CAZymes (the antiSMASH results are available in **Supplementary material** S**10 and Supplementary material S11**). For TH0179, 4 core enzymes (the hybrid NRPS-T1PKS th0179_001, NRPS th179_010, NRPS-like th179_018 and terpene th179_029) plus 60 other SM genes were identified. For TA0020, 7 core enzymes (two

neighboring T1PKS genes in TA0020_003, the NRPS ta20_020, the terpene ta20_013, the fungal RiPP-like ta20_023, the NRPS ta20_030, and the NRPS-like ta20_056) plus 90 other SM genes were linked to the core enzymes. KEGG annotated genes represented approximately 40% of all annotated models for both strains.

CAZymes, transporters and transcription factors and clusters

TH0179 recovered 59 CAZymes, 27 TFs, and 39 TRs. Seventy-one genes presented signal peptides, including 23 CAZymes, one TR, one TF, and two SMGs linked to the NRPS core enzyme from th179_010. Forty-two percent of the annotated CAZymes had predicted substrates, four of which had multiple targets. Regarding CAZyme clustering, in TH0179, three genomic regions did not present any CAZyme, and 5 regions presented only one each. The other 15 selected regions had 2 to 6 CAZymes, among which th179_029 was considered the BAC with the AA2, CE4, GH5, CE4, and AA11 CAZymes; interestingly, all of these were to some extent related to cellulose degradation, with the GH and CE genes considered lignocellulolytic, while the AA redox family is involved in complex substrate degradation. dbCAN3 can also be used to identify signature CAZyme clusters based on genomic localization and homology to known clusters in the database. For TH0179, eight clusters containing CAZyme, TR and TF genes were identified: th179_001 (cluster 1: 2, 1, 1; cluster 2: 2, 3, 1), th179_003 (2, 5, 1), th179_004 (6, 1, 1), th179_0019 (cluster 1: 1, 1, 1; cluster 2: 2, 6, 1), th179_022 (2, 2, 1), and th179_028 (2, 1, 2). (Figure 1)

For TA0020, 143 CAZymes, 62 TFs, and 114 TRs were identified. A total of 131 genes presented signal peptides, including 43 CAZyme, two TF, two TR and five SM genes. Forty percent of TA0020's CAZymes had predicted substrates, with 13 showing more than one target

substrate. Regarding CAZyme clustering, four regions did not present any CAZyme, four regions presented one CAZyme each, and 36 genomic regions presented two or more CAZyme clusters. In the dbCAN3 analysis, 8 regions were identified as CAZyme gene clusters: ta20_001 (1 CAZyme, 1 TR and 1 TF), ta20_005 (3, 2, 2), ta20_012 (2, 5, 1), ta20_031 (1, 1, 3), ta20_040 (3, 4, 1), ta20_043 (1, 2, 2), ta20_49 (1, 6, 1) and ta20_052 (1, 2, 1). (Table 2)

Overall, the TA0020 gene models presented a greater number of annotated CAZymes, TFs and TRs than TH179. The differences in gene number and family classification are shown in Fig 1.

Regarding the SMBGCs, TH0179 recovered four clusters: th179_001 (NRPS-T1PKS hybrid), th179_010 (NRPS-like), th179_018 (NRPS), and th179_029 (terpene). TA0020 recovered seven clusters in 6 genomic regions: ta20_003 (two different T1PKS clusters), ta20_010 (NRPS), ta20_013 (terpene), ta20_023 (Ripp-like), ta20_030 (NRPS), and ta20_056 (NRPS-like).

Expression analysis under cellulose degradation conditions

For all BACs, 69% (893) of the TA0020 genes were expressed. The most highly expressed gene in TA0020, with an impressive TPM value of 239000 on cellulose, presented no annotation. Four genes presented expression levels between 35000 and 85000 TPM, encoding ubiquitin, a protoglobin and a mitochondrial respiratory complex. Seventy-six genes presented between 1000 and 20000 TPM, including 11 CAZyme, 7 TR, 3 TF and two SM genes. A total of 268 genes presented between 1000 and 50 TPM. In addition, 277 genes presented less than 50 TPM. A total of 110 genes presented expression levels between 1000 and 20000 TPM, and 585 presented expression levels between 50 and 1000 TPM. A total of 580 genes presented less than

50 TPM. Differential expression was determined in edgeR. In TA0020, 62 genes showed significant differential expression, 27 on glucose and 35 on cellulose. (Table 3)

The glucose-related DEGs included a GH17 (TA0020_0933), one AA9 (TA0020_1125), a CBM18 (TA0020_1121), three TRs (TA0020_0935, TA0020_0744, TA0020_1208), 6 TFs (TA0020_0369, TA0020_0374, TA0020_0371, TA0020_0358, TA0020_0735, TA0020_1197), an additional biosynthetic SM gene from the T1PKS SMBGC cluster (TA0020_0056), and a direct neighbor gene from the Terpene core-enzyme SMBGC (TA0020_1004). The cellulose DEGs included GH34 (TA0020_0446), GT1 (TA0020_1099), GT20 (TA0020_1261), GH28 (TA0020_0170), three TFs (TA0020_1100, TA0020_0593, TA0020_0023), and two genes from the Ripp-like SMBGC, one with no annotation (TA0020_0494) and one with an alpha-beta hydrolase fold domain (TA0020_0489). Among the DEGs, the functional inference for two on glucose and five on cellulose remained unclear, and these are therefore potential new targets for future characterization.

In regard to the TH0179 BACs, 88% (557) of the genes were expressed, and the most highly expressed gene showed a TPM value of more than 297000 for both conditions. Six genes had expression levels between 90000 and 20000 TPM, including a CAZyme and an SM transporter gene. No DEGs were identified in the recovered regions, most likely due to limited access to the TH0179 genome.

A clustering analysis of all annotated gene models was conducted using the TPM value from StringTie quantification available in the functional annotation table (**Supplementary material 5**). CAZyme gene clustering analysis was conducted to identify major CAZyme players in the studied conditions. (Figure 2)

For TH0179, CAZyme expression levels ranged from 0 to 5600 TPM, and an outlier GH31 presenting more than 47000 TPM under both conditions was identified. Four groups emerged in the clustering analysis: high expression on cellulose (7 genes), high expression on glucose (9 genes), low expression in both conditions (22 genes) and high expression in both conditions (21 genes). Fourteen CAZymes were considered lignocellulolytic. Interestingly, no lignocellulolytic CAZymes were part of the high expression in the cellulose group. Nine CAZymes clustered in the low expression in both conditions group, two mannosidases, TH0179_0598 and TH0179_0599 GH26 (th179_025), clustered in the glycolytic profile group, and the remaining 3, in the cellulose and glucose intermediate expression group.

For TA0020, no outlier was identified, and the CAZyme expression level reached 3500 TPM. Four major groups emerged, including the high expression on cellulose group, in which an 11-gene subgroup showed a strong cellulolytic profile. In the high expression glucose group, a 9-gene subgroup showed a strong glycolytic profile. Regarding lignocellulolytic CAZyme expression, three clustered in the strong cellulolytic profile group, TA0020 _1097 and TA0020_1098 GH5_CBM1 (ta20_012) and one TA0020_0654 GH7 (ta20_001). The other four genes clustered in the high expression on cellulose group. Three lignocellulolytic CAZymes clustered in the glycolytic expression profile: TA0020_0720 GH62 (ta20_041), TA0020_0644 GH30 (ta20_001) and TA0020_1233 GH43 (ta20_050)

Major cellulolytic regions and clusters

For in-depth investigation of the genome regions captured by the BACs, the gene clusters that presented degradative genes were modeled, and the gene position, annotation and/or expression information is shown in Fig. 3 for TA0020 and Fig. 4 for TH0179.

In TA0020, 31 genomic regions had genes with degradative patterns identified via differential expression analysis and in the CAZyme clustering analysis. Thirteen regions were considered cellulolytic, 9 hybrid and 8 glycolytic.

DEGs and the CAZyme expression clustering profiles with cellulolytic/glycolytic profiles are indicators of the relationship to the studied degradative conditions. Three regions presented SMBGCs, and ta20_023, which carries an Ripp-like SMBGC, presented two DEGs, TA0020_0489 S09X alpha-beta hydrolase fold and TA0020_0494, with no functional hit except a hypothetical protein ortholog and the antiSMASH hit. Eight regions presented CAZyme clusters: ta20_001 presented 3 DEGs, TA0020_0645 with an FAD-binding domain, TA0020_0655 G1 peptidase, and TA0020_0656 S-adenosyl-L-methionine-dependent methyltransferase near TA0020_0644 GH7, which presented a cellulolytic profile in the CAZyme clustering analysis. Ta20_012 presented four neighboring genes and two CAZymes identified via CAZyme clustering analysis: TA0020_1097 GH5, TA0020_1098 GH5, TA0020_1099 GT1, and TA0020_1100 TF RTA-like protein. Ta20_052 also presented a CAZyme cluster with DEGs: TA0020_0023, a TF Zn(2)-C6 fungal-type DNA-binding domain, and TA0020_0032, with no functional hit besides a hypothetical protein ortholog identified via COG mapping. (Figure 3)

TH0179 did not present DEGs; therefore, the cellulolytic regions were exclusively identified in the CAZyme expression clustering analysis. Seven CAZymes presented cellulolytic profiles: TH179_0057 GT109 (th179_018), TH179_0458 GH0 (th179_028), TH179_0168 GH2 (th179_009), TH179_0530 GT15 (th179_012), TH179_416 PL38 (th179_001), GH95 TH179_0001 (th179_019), and TH179_0102 (th179_004).

(Figure 4)

Overall, relevant genomic regions with CAZyme-annotated genes were observed. The expression and clustering analysis for TH0179 and Ta0020 helps to illustrate the expression of these genomic regions and their vicinities under cellulose metabolism induction.

Comparative analysis

Alignment to *T. reesei* QM6a was limited for Th0179 and Ta0020, and the alignment rate increased compared to the species reference (Table 4). In comparison to TH0179, TA0020 aligned less well with its own reference than TH0179 did with *T. harzianum* CBS226.95 (Fig. 5). OrthoFinder established orthologous relationships across 30 fungal species. In total, 310808 genes were analyzed, 94.6% of which were assigned to 19020 orthogroups, including 1392 species-specific orthogroups. Forty-eight orthogroups were shared among all strains. A total of 1343 (92.2%) TA0020 genes were assigned to orthogroups, with 37 genes remaining unassigned, and they were distributed in almost all 24 genomic regions. The expression levels of the unassigned genes were between 0 and 300 TPM. Among the TH0179 genes, 596 (94.8%) were assigned to orthogroups, and 33 remained unassigned, all without annotation and expression levels under 450 TPM, and were distributed in 17 genomic regions. In addition, 168 orthogroups were shared among TH0179 and TA0020, including 15 CAZyme, 12 TR, 10 TF and 5 SMB genes.

Discussion

BAC-guided transcriptome analysis was used to mine high-quality genomic regions from TA0020 and TH0179 containing CAZyme targets with biotechnological relevance. This methodology promoted *hybrid* evidence-based and *ab initio* gene prediction, followed by

functional inference and alignment-based expression analysis to identify regions potentially linked to cellulose and glucose degradation. Overall, the main goal was to inspect the genomic and transcriptomic variations and review the genomic context of potential CAZymes and other relevant features related to the carbohydrate metabolism process. A range of clustered CAZymes was identified, including cellulases, hemicellulases, and ligninases. Additionally, important genes, such as secondary metabolite gene clusters, oxidases, proteinases, TFs and TRs, were physically linked to CAZymes in gene clusters. Additionally, genes with high expression levels in degradative conditions were identified and are likely relevant to the process. The exact relationship of clustered genes, how their expression is orchestrated to respond to lignocellulose synergistically, and species-specific particularities affecting the ligninolytic process have yet to be reviewed from a nonmodel species perspective.

Genomic region characterization

In relation to genomic coverage obtained by the BAC approach, the total length for TA0020 was greater than that for TH179, and any numeric differences in the comparison are explained by that limitation. Previous comparative analysis showed that 50-25% of the *Trichoderma* genes are species-specific and not shared with other *Trichoderma* species. Additionally, gene families such as ankyrin, heterokaryon incompatibility domain protein (HET), TF, carbohydrate degradative, and secondary metabolism genes all present high polymorphism rates (Kubicek et al., 2011, 2019; Mukherjee et al., 2013). The intraspecific diversity of the *Harzianum* clade poses a challenge to even modern molecular genetics because this species aggregate has multiple infra groups and great complexity (Druzhinina et al., 2010; Kubicek et al., 2019; Schalamun and Schmoll, 2022), which can partially explain the limited CAZyme transfer.

CAZyme genes and clusters

CAZymes have a central role in carbon substrate degradation, energy consumption, cell wall synthesis, and other accessory activities mediated by substrate availability and other environmental stimuli (Drula et al., 2022).

For the studied species, the retrieved CAZyme genes represent a tiny fraction of all CAZymes (Kubicek et al., 2011; Filho et al., 2017; Rosolen et al., 2023). Whole-genome analysis has shown that *T. reesei* genomes present fewer CAZymes than biocontrol strains (Kubicek et al., 2011; Li et al., 2017; Li and Wang, 2021). This is related to the fact that *T. reesei* is a saprophytic species that has lost biocontrol-related gene families, including important CAZyme genes (Kubicek et al., 2019). A recent whole-genome comparative analysis of *T. harzianum* IOC 3844, TH0179, and TA0020 and *T. reesei* 711 corroborated these differences. For the retrieved genomic regions in the present study, TA0020 showed a diverse set of cellulases, hemicellulases and ligninases despite a lower lignocellulolytic potential, which may be related to plant-*Trichoderma* interactions partially degrading the plant-cell wall, allowing the exchange of nutrients in endophytic species.

Complex substrate degradation, such as that of glucose, cellulose, lignocellulose, and chitin, activates different and complex synergistic reactions with different sets of genes, including CAZymes, TFs, and TRs. The clustered distribution of genes contributing to the same phenotype is not new, but in eukaryotes, genes are expressed individually, and the clustering of metabolic genes in the fungal genome is still unclear (Wisecaver and Rokas, 2015; Slot, 2017; Rokas et al., 2018). Regarding the clustered distribution of CAZymes, overall, more than half of the CAZyme-containing BACs presented multiple CAZymes with important regulators,

transporters, and other genes with a variety of functions. Careful inspection of such clusters must be conducted to understand the coregulatory patterns and the synergistic actions among clustered genes to clarify the lignocellulolytic activities of *Trichoderma* ssp.

Expression measurement and DEG identification

A prevalidated RNA-Seq (Almeida et al., 2021) dataset was used as a source to identify regions/genes potentially linked to lignocellulose degradation. A previous study characterized the global transcriptome through transcriptomic profiling, revealing different strategies for carbon utilization (Almeida et al., 2021). As transcriptomic, genomic, and phenotypic variation is common even at the strain level, such differences must be carefully studied to understand their role in the biodegradative process (Hewedy et al., 2020; Almeida et al., 2021; Pachauri et al., 2023; Rosolen et al., 2023). In TA0020, 31 genes showed expression levels greater than 5000 TPM. In the differential expression analysis, only a few genes showed significant differential expression and were likely regulated to respond to the studied conditions. The lack of DEGs in TH0179 can likely be attributed to the limited access to its genome.

The complex transcriptomic profile of TA0020, observed in the previous *de novo* transcriptome analysis, was captured in this reference-based transcriptome assembly and differential expression analysis. The lower lignocellulolytic potential contrasts with the complex transcriptome profile, likely because *T. atroviride* may not be as specialized in lignocellulose utilization as the saprophytic workhorse *T. reesei* QM6a or the *T. harzianum* IOC 3844 and *T. harzianum* CBMAI-0179 strains, which show a higher efficiency in crystalline cellulose degradation (Horta et al., 2018; Almeida et al., 2021; Li and Wang, 2021).

Comparative analysis

The BAC alignment with the reference genome and orthology establishment helped in the assessment of the complexity of *Trichoderma* at the strain level. *T. reesei* QM6a is considered a model *Trichoderma* species that is phylogenetically distant from other biocontrol species (Kubicek et al., 2019). As expected, the BAC correspondence decreased with phylogenetic distance. Based on the alignment to its own species reference, it is possible to conclude that TA0020 may be more distant from *T. atroviride* IMI206040 than TH0179 is from *T. harzianum* CBS226.95. Additionally, for all three comparisons, the TA0020 alignments showed a greater complexity, with the same BAC aligning to multiple contigs/chromosomes in the references. In the orthology analysis, only 33 TH0179 genes were unassigned to orthogroups, while 37 genes for TA0020 remained unassigned. In a previous molecular phylogenetic analysis of rpb2, *T. atroviride* species belonged to a basal branch that diverged early from other species, most likely maintaining characteristics from the common ancestor (Kubicek et al., 2019).

The opportunistic lifestyle of *Trichoderma* is characterized by fast evolution rates that reshape the genomes, allowing them to colonize potential niches (Kubicek et al., 2019). Their remarkable capacities at the intraspecific level are greatly influenced by lateral gene transfer and other complex evolutionary processes contributing to genomic variations, even those observed in closely related species. Interestingly, carbohydrate metabolism and biocontrol gene families seem to be especially influenced by evolutionary processes (Kubicek et al., 2011; Kubicek et al., 2019; Hewedy et al., 2020; Sood et al., 2020; Schalamun and Schmoll, 2022; Tyśkiewicz et al., 2022; Guzmán-Guzmán et al., 2023).

Conclusions

The main goal was to recover CAZyme-containing regions and study their linear distribution, genomic context, and relationship to degradative conditions to identify potential lignocellulolytic CAZyme clusters, aiming to address the differential lignocellulolytic potential previously observed (Almeida et al., 2021). Functional inference combined with transcript expression and comparative analysis provided highly valuable information regarding the annotated genes; thus, it was possible to identify gaps in *Trichoderma* functional research and strain-specific characteristics. We recovered multiple CAZyme-containing regions rich in TF- and TR-flanking genes. Expression analysis helped to identify regions positively regulated in biomass likely involved in the lignocellulolytic activities of both strains. The results contribute to a genomic perspective of targeted CAZyme regions in nonmodel *Trichoderma* species, providing high-quality genomic knowledge of novel *Trichoderma* strains with different biocontrol and lignocellulolytic potentials.

Acknowledgments

Funding

Financial support for this work was provided by the São Paulo Research Foundation (FAPESP - Process number 2015/09202-0 and 2018/19660-4) and the Coordination for the Improvement of Higher Education Personnel (CAPES, Computational Biology Program - Process number 88882.160095/2013-01). PHCA received a PhD fellowship from CAPES (88887.612254/2021-00), RRR received a PhD fellowship from CAPES (88887.482201/2020-00) and FAPESP (2020/13420-1), MACH received a postdoctoral fellowship from FAPESP (2020/10536-9), and

APS received a research fellowship from the Brazilian National Council for Technological and Scientific Development (CNPq- Process number 312777/2018-3).

Role of the funding source

CAPES and FAPESP were responsible for funding scholarships and the research

Declaration of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as potential conflicts of interest.

Declaration of generative AI and AI-assisted technologies in the writing process

No AI-based technology was used in the writing process.

Data availability

All data generated or analyzed in this study are included in this published article (and its supplementary information files). The assembled genomic regions were deposited in the NCBI database via BankIt and can be accessed under BioProject number XXXXXXXX and BioSample numbers x (TH0179) and y (TA0020). Raw sequencing data deposited on NCBI SRA servers.

Authors' contributions

PHC: writing—original draft, methodology, formal analysis, and visualization.

RRR: methodology and writing—review & editing.

MACH: methodology and writing—review & editing.

JSM: methodology and resources.

CCS: methodology and resources.

DAS: methodology and resources.

APS: conceptualization, supervision, review & editing, and funding acquisition.

References

Almeida, D.A., Horta, M.A.C., Ferreira Filho, J.A., Murad, N.F., de Souza, A.P., 2021. The synergistic actions of hydrolytic genes reveal the mechanism of trichoderma harzianum for cellulose degradation. J. Biotechnol. 334, 1–10. https://doi.org/10.1016/j.jbiotec.2021.05.001.

Andlar, M., Rezić, T., Marđetko, N., Kracher, D., Ludwig, R., Šantek, B., 2018. Lignocellulose degradation: An overview of fungi and fungal enzymes involved in lignocellulose degradation. Eng. Life Sci. 18, 768–778. https://doi.org/10.1002/elsc.201800039.

Baroncelli, R., Piaggeschi, G., Fiorini, L., Bertolini, E., Zapparata, A., Pè, M.E., Sarrocco, S., Vannacci, G., 2015. Draft whole-genome sequence of the biocontrol agent trichoderma harzianum T6776. Genome Announc. 3, e00647–15. https://doi.org/10.1128/genomeA.00647-15.

Bischof, R. H., Ramoni, J., Seiboth, B. 2016. Cellulases and beyond: the first 70 years of the enzyme producer *Trichoderma reesei*. Microbial cell factories v15, issue 1. doi: 10.1186/s12934-016-0507-6

Blin, K., Shaw, S., Kloosterman, A.M., Charlop-Powers, Z., van Wezel, G.P., Medema, M.H., Weber, T., 2021. antiSMASH 6.0: Improving cluster detection and comparison capabilities. Nucleic Acids Res. 49, W29–W35. https://doi.org/10.1093/nar/gkab335.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–20. https://doi.org/10.1093/bioinformatics/btu170.

Crucello, A., Sforça, D.A., Horta, M.A., dos Santos, C.A., Viana, A.J., Beloti, L.L., de Toledo, M.A., Vincentz, M., Kuroshu, R.M., de Souza, A.P., 2015. Analysis of genomic regions

of trichoderma harzianum IOC-3844 related to biomass degradation. PLoS One 10, e0122122. https://doi.org/10.1371/journal.pone.0122122.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21. https://doi.org/10.1093/bioinformatics/bts635.

Drula, E., Garron, M.L., Dogan, S., Lombard, V., Henrissat, B., Terrapon, N., 2022. The carbohydrate-active enzyme database: Functions and literature. Nucleic Acids Res. 50, D571–D577. https://doi.org/10.1093/nar/gkab1045.

Druzhinina, I.S., Kubicek, C.P., Komoń-Zelazowska, M., Mulaw, T.B., Bissett, J., 2010. The trichoderma harzianum demon: Complex speciation history resulting in coexistence of hypothetical biological species, recent agamospecies and numerous relict lineages. BMC Evol. Biol. 10, 94. https://doi.org/10.1186/1471-2148-10-94.

Druzhinina, I. S. ;Chenthamara, K., Zhang, J., Atanasova, L.Yang, D., Miao, Y., Rahimi, M.J., Grujic, M., Cai, F., Pourmehdi, S., Salim, K. A., Pretzer, C., Kopchinskiy, A. G., Henrissat, B., Kuo, A., Hundley, H., Wang, M., Aerts, A., Salamov, A., Lipzen, A., LaButti, K., Barry, K., Grigoriev, I. V., Shen, Q., Kubicek, C. P. Massive lateral transfer of genes encoding plant cell wall-degrading enzymes to the mycoparasitic fungus Trichoderma from its plant-associated hosts. PLOS genetics. doi: 10.1371/journal.pgen.1007322

Emms, D. M, Kelly, S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biology. vol. 20, issue 1. doi: 10.1186/s13059-019-1832-y

Filho, J.A.F., Horta, M.A.C., Beloti, L.L., Dos Santos, C.A., de Souza, A.P., 2017. Carbohydrate-active enzymes in trichoderma harzianum: A bioinformatic analysis

bioprospecting for key enzymes for the biofuels industry. BMC Genom. 18, 779. https://doi.org/10.1186/s12864-017-4181-9.

Filho, J.A.F., Horta, M.A.C., dos Santos, C.A., Almeida, D.A., Murad, N.F., Mendes, J.S., Sforça, D.A., Silva, C.B.C., Crucello, A., de Souza, A.P., 2020. Integrative genomic analysis of the bioprospection of regulators and accessory enzymes associated with cellulose degradation in a filamentous fungus (Trichoderma harzianum). BMC Genom. 21, 757. https://doi.org/10.1186/s12864-020-07158-w.

Galagan, J.E., Calvo, S.E., Borkovich, K.A., Selker, E.U., Read, N.D., Jaffe, D., FitzHugh, W., Ma, L.-J., Smirnov, S., Purcell, S., Rehman, B., Elkins, T., Engels, R., Wang, S., Nielsen, C.B., Butler, J., Endrizzi, M., Qui, D., Ianakiev, P., Bell-Pedersen, D., Nelson, M.A., Werner-Washburne, M., Selitrennikoff, C.P., Kinsey, J.A., Braun, E.L., Zelter, A., Schulte, U., Kothe, G.O., Jedd, G., Mewes, W., Staben, C., Marcotte, E., Greenberg, D., Roy, A., Foley, K., Naylor, J., Stange-Thomann, N., Barrett, R., Gnerre, S., Kamal, M., Kamvysselis, M., Mauceli, E., Bielke, C., Rudd, S., Frishman, D., Krystofova, S., Rasmussen, C., Metzenberg, R.L., Perkins, D.D., Kroken, S., Cogoni, C., Macino, G., Catcheside, D., Li, W., Pratt, R.J., Osmani, S.A., DeSouza, C.P.C., Glass, L., Orbach, M.J., Berglund, J.A., Voelker, R., Yarden, O., Plamann, M., Seiler, S., Dunlap, J., Radford, A., Aramayo, R., Natvig, D.O., Alex, L.A., Mannhaupt, G., Ebbole, D.J., Freitag, M., Paulsen, I., Sachs, M.S., Lander, E.S., Nusbaum, C., Birren, B., 2003. The genome sequence of the filamentous fungus neurospora crassa. Nature 422, 859–868. https://doi.org/10.1038/nature01554.

Galaxy Community, 2022. The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. Nucleic Acids Res. 50, W345–W351. https://doi.org/10.1093/nar/gkac247.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H., Oliver, S.G., 1996. Life with 6000 genes. Science 274, 546, 563–7. https://doi.org/10.1126/science.274.5287.546.

Grigoriev, I.V., Nikitin, R., Haridas, S., Kuo, A., Ohm, R., Otillar, R., Riley, R., Salamov, A., Zhao, X., Korzeniewski, F., Smirnova, T., Nordberg, H., Dubchak, I., Shabalov, I., 2014. MycoCosm portal: Gearing up for 1000 fungal genomes. Nucleic Acids Res. 42, D699–D704. https://doi.org/10.1093/nar/gkt1183.

Guzmán-Guzmán, P., Kumar, A., de Los Santos-Villalobos, S., Parra-Cota, F.I., Orozco-Mosqueda, M.D.C., Fadiji, A.E., Hyder, S., Babalola, O.O., Santoyo, G., 2023. Trichoderma species: Our best fungal allies in the biocontrol of plant diseases-a review. Plants (Basel) 12, 432. https://doi.org/10.3390/plants12030432.

Hewedy, O.A., Abdel-Lateif, K.S., Bakr, R.A., 2020. Genetic diversity and biocontrol efficacy of indigenous Trichoderma isolates against Fusarium wilt of pepper. J. Basic Microbiol. 60, 126–135. https://doi.org/10.1002/jobm.201900493.

Holt, C., Yandell, M., 2011. MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinform. 12, 491. https://doi.org/10.1186/1471-2105-12-491.

Horta, M.A.C., Filho, J.A.F., Murad, N.F., de Oliveira Santos, E., dos Santos, C.A., Mendes, J.S., Brandão, M.M., Azzoni, S.F., de Souza, A.P., 2018. Network of proteins, enzymes

and genes linked to biomass degradation shared by trichoderma species. Sci. Rep. 8, 1341. https://doi.org/10.1038/s41598-018-19671-w.

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S.K., Cook, H., Mende, D.R., Letunic, I., Rattei, T., Jensen, L.J., von Mering, C., Bork, P., 2019. eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 47, D309–D314. https://doi.org/10.1093/nar/gky1085.

Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A.F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.Y., Lopez, R., Hunter, S., 2014. InterProScan 5: Genome-scale protein function classification. Bioinformatics 30, 1236–1240. https://doi.org/10.1093/bioinformatics/btu031.

Kanehisa, M., Goto, S., 2000. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28, 27–30. https://doi.org/10.1093/nar/28.1.27.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., Tanabe, M., 2016. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 44, D457–D462. https://doi.org/10.1093/nar/gkv1070.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., Phillippy, A.M., 2017. Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27, 722–736. https://doi.org/10.1101/gr.215087.116.

Kubicek, C.P., Herrera-Estrella, A., Seidl-Seiboth, V., Martinez, D.A., Druzhinina, I.S., Thon, M., Zeilinger, S., Casas-Flores, S., Horwitz, B.A., Mukherjee, P.K., Mukherjee, M., Kredics, L., Alcaraz, L.D., Aerts, A., Antal, Z., Atanasova, L., Cervantes-Badillo, M.G.,

Challacombe, J., Chertkov, O., McCluskey, K., Coulpier, F., Deshpande, N., von Döhren, H., Ebbole, D.J., Esquivel-Naranjo, E.U., Fekete, E., Flipphi, M., Glaser, F., Gómez-Rodríguez, E.Y., Gruber, S., Han, C., Henrissat, B., Hermosa, R., Hernández-Oñate, M., Karaffa, L., Kosti, I., Le Crom, S., Lindquist, E., Lucas, S., Lübeck, M., Lübeck, P.S., Margeot, A., Metz, B., Misra, M., Nevalainen, H., Omann, M., Packer, N., Perrone, G., Uresti-Rivera, E.E., Salamov, A., Schmoll, M., Seiboth, B., Shapiro, H., Sukno, S., Tamayo-Ramos, J.A., Tisch, D., Wiest, A., Wilkinson, H.H., Zhang, M., Coutinho, P.M., Kenerley, C.M., Monte, E., Baker, S.E., Grigoriev, I.V., 2011. Comparative genome sequence analysis underscores mycoparasitism as the ancestral life style of trichoderma. Genome Biol. 12, R40. https://doi.org/10.1186/gb-2011-12-4-r40.

Kubicek, C.P., Steindorff, A.S., Chenthamara, K., Manganiello, G., Henrissat, B., Zhang, J., Cai, F., Kopchinskiy, A.G., Kubicek, E.M., Kuo, A., Baroncelli, R., Sarrocco, S., Noronha, E.F., Vannacci, G., Shen, Q., Grigoriev, I.V., Druzhinina, I.S., 2019. Evolution and comparative genomics of the most common trichoderma species. BMC Genom. 20, 485. https://doi.org/10.1186/s12864-019-5680-7.

Li, W.-C., Wang, T.-F., 2021. PacBio long-read sequencing, assembly, and funannotate reannotation of the complete genome of trichoderma reesei QM6a, in: Mach-Aigner, A.R., Martzy, R. (Eds.), Trichoderma Reesei: Methods and Protocols. Springer US, New York, NY, pp. 311–329.

Li, W.C., Huang, C.H., Chen, C.L., Chuang, Y.C., Tung, S.Y., Wang, T.F., 2017. Trichoderma reesei complete genome sequence, repeat-induced point mutation, and partitioning of CAZyme gene clusters. Biotechnol. Biofuels 10, 170. https://doi.org/10.1186/s13068-017-0825-x.

Mancini, M.C., Cardoso-Silva, C.B., Sforça, D.A., Pereira de Souza, A., 2018. "Targeted sequencing by gene synteny," a new strategy for polyploid species: Sequencing and physical structure of a complex sugarcane region. Front. Plant Sci. 9, 397. https://doi.org/10.3389/fpls.2018.00397.

Marçais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L., Zimin, A., 2018. MUMmer4: A fast and versatile genome alignment system. PLoS Comput. Biol. 14, e1005944. https://doi.org/10.1371/journal.pcbi.1005944.

Mukherjee, P.K., Horwitz, B.A., Herrera-Estrella, A., Schmoll, M., Kenerley, C.M., 2013. Trichoderma research in the genome era. Annu. Rev. Phytopathol. 51, 105–9. https://doi.org/10.1146/annurev-phyto-082712-102353.

Nascimento, V.C., Rodrigues-Santos, K.C., Carvalho-Alencar, K.L., Castro, M.B., Kruger, R.H., Lopes, F.A.C., 2022. Trichoderma: Biological control efficiency and perspectives for the Brazilian midwest states and Tocantins. Braz. J. Biol. 82, e260161. https://doi.org/10.1590/1519-6984.260161.

Ohm, R.A., Riley, R., Salamov, A., Min, B., Choi, I.G., Grigoriev, I.V., 2014. Genomics of wood-degrading fungi. Fungal Genet. Biol. 72, 82–90. https://doi.org/10.1016/j.fgb.2014.05.001.

Pachauri, S., Zaid, R., Sherkhane, P.D., Easa, J., Viterbo, A., Chet, I., Horwitz, B.A., Mukherjee, P.K., 2023. Comparative phenotypic, genomic, and transcriptomic analyses of two contrasting strains of the plant beneficial fungus trichoderma virens. Microbiol. Spectr. 11, e0302422. https://doi.org/10.1128/spectrum.03024-22.

Palmer, M.J., Stajich, J., 2020. Funannotate v1.8.1: Eukaryotic genome annotation (v1.8.1). Zenodo. https://doi.org/10.5281/zenodo.4054262.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T., Salzberg, S.L., 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. 33, 290–295. https://doi.org/10.1038/nbt.3122.

Peterson, D.G., Tomkins, J.P., Frisch, D.A., Wing, R.A., Paterson, A.H., 2000. Construction of plant bacterial artificial chromosome (BAC) libraries: An illustrated guide. J. Agric. Genom. 5, 1–100. https://doi.org/10.1007/978-94-011-4217-5_1.

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., Lopez, R., 2005. InterProScan: Protein domains identifier. Nucleic Acids Res. 33, W116–W120. https://doi.org/10.1093/nar/gki442.

Riley, R., Salamov, A.A., Brown, D.W., Nagy, L.G., Floudas, D., Held, B.W., Levasseur, A., Lombard, V., Morin, E., Otillar, R., Lindquist, E.A., Sun, H., LaButti, K.M., Schmutz, J., Jabbour, D., Luo, H., Baker, S.E., Pisabarro, A.G., Walton, J.D., Blanchette, R.A., Henrissat, B., Martin, F., Cullen, D., Hibbett, D.S., Grigoriev, I.V., 2014. Extensive sampling of basidiomycete genomes demonstrates inadequacy of the white-rot/brown-rot paradigm for wood decay fungi. Proc. Natl. Acad. Sci. U. S. A. 111, 9923–8. https://doi.org/10.1073/pnas.1400592111.

Robinson, M.D., McCarthy, D.J., Smyth, G.K., 2010. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–40. https://doi.org/10.1093/bioinformatics/btp616.

Rokas, A., Wisecaver, J.H., Lind, A.L., 2018. The birth, evolution and death of metabolic gene clusters in fungi. Nat. Rev. Microbiol. 16, 731–744. https://doi.org/10.1038/s41579-018-0075-3.

Rosolen, R.R., Aono, A.H., Almeida, D.A., Ferreira Filho, J.A., Horta, M.A.C., De Souza, A.P., 2022. Network analysis reveals different cellulose degradation strategies across trichoderma harzianum strains associated with XYR1 and CRE1. Front. Genet. 13, 807243. https://doi.org/10.3389/fgene.2022.807243.

Rosolen, R.R., Horta, M.A.C., de Azevedo, P.H.C., da Silva, C.C., Sforca, D.A., Goldman, G.H., de Souza, A.P., 2023. Whole-genome sequencing and comparative genomic analysis of potential biotechnological strains of trichoderma harzianum, trichoderma atroviride, and trichoderma reesei. Mol. Genet. Genom. 298, 735–754. https://doi.org/10.1007/s00438-023-02013-5.

Schalamun, M., Schmoll, M., 2022. Trichoderma – genomes and genomics as treasure troves for research toward biology, biotechnology and agriculture. Front. Fungal Biol. 3, 1002161. https://doi.org/10.3389/ffunb.2022.1002161.

Slot, J.C., 2017. Fungal gene cluster diversity and evolution. Adv. Genet. 100, 141–178. https://doi.org/10.1016/bs.adgen.2017.09.005.

Sood, M., Kapoor, D., Kumar, V., Sheteiwy, M.S., Ramakrishnan, M., Landi, M., Araniti, F., Sharma, A., 2020. Trichoderma: The "secrets" of a multitalented biocontrol agent. Plants (Basel) 9, 762. https://doi.org/10.3390/plants9060762.

Srivastava, M. and Shahid, M. 2014. Trichoderma Genome to Genomics: A Review. Journal of Data Mining in Genomics & Proteomics. doi: 10.4172/2153-0602.1000162Stracquadanio, C., Quiles, J.M., Meca, G., Cacciola, S.O., 2020. Antifungal activity of bioactive metabolites produced by trichoderma asperellum and trichoderma atroviride in liquid medium. J. Fungi (Basel) 6, 263. https://doi.org/10.3390/jof6040263.

Trivedi, U.H., Cézard, T., Bridgett, S., Montazam, A., Nichols, J., Blaxter, M., Gharbi, K., 2014. Quality control of next-generation sequencing data without a reference. Front. Genet. 5, 111. https://doi.org/10.3389/fgene.2014.00111.

Tyśkiewicz, R., Nowak, A., Ozimek, E., Jaroszuk-Ściseł, J., 2022. Trichoderma: The current status of its application in agriculture for the biocontrol of fungal phytopathogens and stimulation of plant growth. Int. J. Mol. Sci. 23, 2329. https://doi.org/10.3390/ijms23042329.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M., Rozen, S.G., 2012. Primer3--new capabilities and interfaces. Nucleic Acids Res. 40, e115. https://doi.org/10.1093/nar/gks596.

Visendi, P., Berkman, P.J., Hayashi, S., Golicz, A.A., Bayer, P.E., Ruperao, P., Hurgobin, B., Montenegro, J., Chan, C.K., Staňková, H., Batley, J., Šimková, H., Doležel, J., Edwards, D., 2016. An efficient approach to BAC based assembly of complex genomes. Plant Methods 12, 2. https://doi.org/10.1186/s13007-016-0107-9.

Waterston, R.H., Lander, E.S., Sulston, J.E., 2002. On the sequencing of the human genome. Proc. Natl. Acad. Sci. U. S. A. 99, 3712–6. https://doi.org/10.1073/pnas.042692499.

Wisecaver, J.H., Rokas, A., 2015. Fungal metabolic gene clusters-caravans traveling across genomes and environments. Front. Microbiol. 6, 161. https://doi.org/10.3389/fmicb.2015.00161.

Xu, M., Wang, Y., Zhao, Z., Gao, G., Huang, S.X., Kang, Q., He, X., Lin, S., Pang, X., Deng, Z., Tao, M., 2016. Functional genome mining for metabolites encoded by large gene clusters through heterologous expression of a whole-genome bacterial artificial chromosome

library in streptomyces spp. Appl. Environ. Microbiol. 82, 5795–805. https://doi.org/10.1128/aem.01383-16.

Zhang, Z., Running, K.L.D., Seneviratne, S., Haugrud, A.R.P., Szabo-Hever, A., Shi, G., Brueggeman, R., Xu, S.S., Friesen, T.L., Faris, J.D., 2021. A protein kinase-major sperm protein gene hijacked by a necrotrophic fungal pathogen triggers disease susceptibility in wheat. Plant J. 106, 720–732. https://doi.org/10.1111/tpj.15194.

Zheng, J., Hu, B., Zhang, X., Ge, Q., Yan, Y., Akresi, J., Piyush, V., Huang, L., Yin, Y., 2023. dbCAN-seq update: CAZyme gene clusters and substrates in microbiomes. Nucleic Acids Res. 51, D557–D563. https://doi.org/10.1093/nar/gkac1068.

Figure captions

Fig. 1 (**A)** direct comparison of relevant gene categories between TA0020 and TH0179**. B** Comparison of lignocellulolytic CAZyme families between TA0020 and TH0179.

Fig. 2 TA0020 and TH0179 CAZyme clustering analysis comparing cellulose and glucose degradative conditions. * Represents lignocellulolytic CAZyme families from Andlar et al. 2022, and the red + identifies the expression outlier from TH0179.

Fig. 3 TA0020 IGV visualization of major degradative regions identified via CAZyme expression clustering analysis and differential expression analysis. The red * indicates the CAZyme identified by clustering analysis.

Fig. 4 TH0179 IGV visualization of major cellulolytic genomic regions identified via CAZyme expression clustering analysis. The red * indicates the CAZyme identified by clustering analysis. The red + indicates the outlier CAZyme.

Fig. 5 Alignments visualized using a MUMmer dot plot with the parameters Minmatch 50 bp and match length 500 bp against *T. reesei* QM6a, *T. harzianum* CBS22695 and *T. atroviride* IMI 206040. **A** TH0179 x *T. reesei* QM6a. **B** TA0020 x *T. reesei* QM6a**. C** TH0179 **x** *T. atroviride* IMI206040. **D** TA0020 **x** *T. atroviride* IMI206040**. E** TH0179 **x** *T. harzianum* CBS226.95 **F** TA0020 **x** *T. harzianum* CBS226.95.

Tables

Table 1. BAC characterization and functional annotation overview of the evaluated strains

	TH0179	TA0020
Sequenced BAC	24	44
Sequenced length (bp)	1955378	4721491
Mean BAC length (bp)	81474.08	107306.61
GC (%)	49.64%	49.94%
Predicted gene models	629	1280
Average gene length (bp)	1704.13	1953.11
Multiexon transcripts	615	1267
Complete CDS	589	1207
Interpro	492	1054
Gene ontology	404	889
Eggnog	542	1138
Pfam	437	948
Cazy	59	142
Signalp	71	131
Merops	20	60
Kegg	181	432
Secondary metabolism core enzyme	4	7

Table 2. CAZyme class distribution

Class	TH0179	%	TA0020	%
CBM	2 + 4*	3,4	6 + 20*	4,2
AA	5	8,5	20	14,1
GH	38	64,4	79	55,7
GT	8	13,6	27	19
CE	5	8,5	9	6,3
PL	1	1,7	2	1,4
Total	59	100	142	100

*Represents annotated CAZyme with additional CBM modules

Table 3. TA0020 DEG annotation

Gene	Log FC	BAC	Category	EC Number	CAZy	CBM	Interpro	COG
TA0020_0936	2,420572022	ta20_027						
TA0020_1258	1,876823142	ta20_038						
TA0020_0309	-1,77480218	ta20_011						
TA0020_1222	-1,96635772	ta20_044						
TA0020_0472	-2,12291573	ta20_028					IPR000719 Protein kinase domain;IPR011009 Protein kinase-like domain superfamily;IPR045269 Serine/threonine-protein kinase Atg1-like	
TA0020_0114	-2,13181176	ta20_040					IPR002110 Ankyrin repeat;IPR036770 Ankyrin repeat-containing domain superfamily	
TA0020_1038	-2,32101001	ta20_025						
TA0020_1175	-2,33431055	ta20_024						
TA0020_0494	-3,82704914	ta20_023	SM					
TA0020_1016	1,775772167	ta20_013		3.6.4.13			IPR000629 ATP-dependent RNA helicase DEAD-box, conserved site;IPR001650 Helicase, C-terminal;IPR011545 DEAD/DEAH box helicase domain;IPR014001 Helicase superfamily 1/2, ATP-binding domain;IPR014014 RNA helicase, DEAD-box type, Q motif;IPR027417 P-loop containing nucleoside triphosphate hydrolase	A:(A) RNA processing and modification
TA0020_0685	-1,85751759	ta20_053					IPR001680 WD40 repeat;IPR015943 WD40/YVTN repeat-like-containing domain superfamily;IPR019775 WD40 repeat, conserved site;IPR020472 G-protein beta WD-40 repeat;IPR036322 WD40-repeat-containing domain superfamily	A:(A) RNA processing and modification
TA0020_0440	-1,86472992	ta20_018		3.5.4.34			IPR002466 Adenosine deaminase/editase;IPR042935 tRNA-specific adenosine deaminase 1	A:(A) RNA processing and modification
TA0020_0358	1,604914287	ta20_005	TF				IPR007219 Transcription factor domain, fungi;IPR011598 Myc-type, basic helix-loop-helix (bHLH) domain;IPR036638 Helix-loop-helix DNA-binding domain superfamily	B:(B) Chromatin structure and dynamics

TA0020_0735	1,567929311	ta20_020	TF				IPR001138 Zn(2)-C6 fungal-type DNA-binding domain;IPR007219 Transcription factor domain, fungi;IPR036864 Zn(2)-C6 fungal-type DNA-binding domain superfamily	B:(B) Chromatin structure and dynamics
TA0020_0802	1,947384641	ta20_054					IPR002871 NIF system FeS cluster assembly, NifU, N-terminal;IPR011339 ISC system FeS cluster assembly, IscU scaffold	C:(C) Energy production and conversion
TA0020_0646	1,531567235	ta20_001		1.14.13.1			IPR002938 FAD-binding domain;IPR036188 FAD/NAD(P)-binding domain superfamily	C:(C) Energy production and conversion
TA0020_0195	1,582734366	ta20_030					IPR015942 Asp/Glu/hydantoin racemase	E:(E) Amino acid transport and metabolism
TA0020_0056	1,511981353	ta20_003	SM	1.2.1.5			IPR015590 Aldehyde dehydrogenase domain;IPR016161 Aldehyde/histidinol dehydrogenase;IPR016162 Aldehyde dehydrogenase, N-terminal;IPR016163 Aldehyde dehydrogenase, C-terminal;IPR029510 Aldehyde dehydrogenase, glutamic acid active site	E:(E) Amino acid transport and metabolism
TA0020_1115	-1,71029179	ta20_021		5.4.99.5			IPR008238 Chorismate mutase, AroQ class, eukaryotic type;IPR036263 Chorismate mutase type II superfamily;IPR037039 Chorismate mutase, AroQ class superfamily, eukaryotic	E:(E) Amino acid transport and metabolism
TA0020_0935	2,154070617	ta20_027	TR				IPR009262 Solute carrier family 35 member SLC35F1/F2/F6	E:(E) Amino acid transport and metabolism;G:(G) Carbohydrate transport and metabolism
TA0020_0744	1,752647842	ta20_020	TR				IPR009262 Solute carrier family 35 member SLC35F1/F2/F6;IPR025016 Protein of unknown function DUF3955	E:(E) Amino acid transport and metabolism;G:(G) Carbohydrate transport and metabolism
TA0020_0933	2,012232363	ta20_027	CAZy	3.2.1.39	GH17		IPR017853 Glycoside hydrolase superfamily	G:(G) Carbohydrate transport and metabolism
TA0020_0446	-1,53824809	ta20_028	CAZy	3.2.1.37	GH34	CBM91	IPR006710 Glycoside hydrolase, family 43;IPR013320 Concanavalin A-like lectin/glucanase domain superfamily;IPR023296 Glycosyl hydrolase, five-bladed beta-propellor domain superfamily;IPR041542 Beta-xylosidase, C-terminal Concanavalin A-like domain	G:(G) Carbohydrate transport and metabolism
TA0020_1193	-1,75543814	ta20_024		4.1.3.16			IPR002220 DapA-like;IPR013785 Aldolase-type TIM barrel	H:(H) Coenzyme transport and metabolism
TA0020_0127	-1,53092393	ta20_040		3.2.1.4			IPR000873 AMP-dependent synthetase/ligase;IPR020845 AMP-binding, conserved site;IPR025110 AMP-binding enzyme, C-terminal domain;IPR042099 ANL, N-terminal domain;IPR045851 AMP-binding enzyme, C-terminal domain superfamily	I:(I) Lipid transport and metabolism

TA0020_1218	-2,85155679	ta20_044			IPR001680 WD40 repeat;IPR015943 WD40/YVTN repeat-like-containing domain superfamily;IPR036322 WD40-repeat-containing domain superfamily	I:(I) Lipid transport and metabolism
TA0020_1228	-1,78246169	ta20_044		3.5.1.4	IPR020556 Amidase, conserved site;IPR023631 Amidase signature domain;IPR036928 Amidase signature (AS) superfamily	I:(I) Lipid transport and metabolism;J:(J) Translation, ribosomal structure and biogenesis;T:(T) Signal transduction mechanisms
TA0020_0765	1,784941208	ta20_049			IPR000915 60S ribosomal protein L6E;IPR008991 Translation protein SH3-like domain superfamily;IPR014722 Ribosomal protein L2, domain 2;IPR041997 Ribosomal Protein L6, KOW domain	J:(J) Translation, ribosomal structure and biogenesis
TA0020_0362	1,708598223	ta20_005		3.4.11.18	IPR000994 Peptidase M24;IPR001714 Peptidase M24, methionine aminopeptidase;IPR002468 Peptidase M24A, methionine aminopeptidase, subfamily 2;IPR036005 Creatinase/aminopeptidase-like;IPR036388 Winged helix-like DNA-binding domain superfamily;IPR036390 Winged helix DNA-binding domain superfamily	J:(J) Translation, ribosomal structure and biogenesis
TA0020_1197	1,544206831	ta20_024	TF		IPR003163 Transcription regulator HTH, APSES-type DNA-binding domain;IPR018004 KilA, N-terminal/APSES-type HTH, DNA-binding;IPR029790 EFG1/Phd1/StuA family;IPR036887 HTH APSES-type DNA-binding domain superfamily	K:(K) Transcription
TA0020_0441	-1,50166891	ta20_018			IPR002715 Nascent polypeptide-associated complex NAC domain;IPR038187 NAC A/B domain superfamily;IPR039370 Transcription factor BTF3	K:(K) Transcription
TA0020_1114	-2,1084915	ta20_021			IPR025602 BCP1 family	K:(K) Transcription
TA0020_0371	2,544153435	ta20_056	TF		IPR001138 Zn(2)-C6 fungal-type DNA-binding domain;IPR007219 Transcription factor domain, fungi;IPR036864 Zn(2)-C6 fungal-type DNA-binding domain superfamily	L:(L) Replication, recombination and repair
TA0020_0438	1,65026467	ta20_018			IPR006597 Sel1-like repeat;IPR011990 Tetratricopeptide-like helical domain superfamily	M:(M) Cell wall/membrane/envelope biogenesis;O:(O) Posttranslational modification, protein turnover, chaperones;T:(T) Signal transduction mechanisms
TA0020_0370	2,900224701	ta20_056			IPR000626 Ubiquitin-like domain;IPR000938 CAP Gly-rich domain;IPR029071 Ubiquitin-like domain superfamily;IPR036859 CAP Gly-rich domain superfamily;IPR045172 Tubulin-folding cofactor B, N-terminal ubiquitin-like domain	O:(O) Posttranslational modification, protein turnover, chaperones

TA0020_0579	1,780747353	ta20_034				IPR001623 DnaJ domain;IPR003604 Matrin/U1-C-like, C2H2-type zinc finger;IPR013087 Zinc finger C2H2-type;IPR018253 DnaJ domain, conserved site;IPR022755 Zinc finger, double-stranded RNA binding;IPR036236 Zinc finger C2H2 superfamily;IPR036869 Chaperone J-domain superfamily	O:(O) Posttranslational modification, protein turnover, chaperones
TA0020_1208	1,644488985	ta20_037	TR			IPR003959 ATPase, AAA-type, core;IPR027417 P-loop containing nucleoside triphosphate hydrolase	O:(O) Posttranslational modification, protein turnover, chaperones
TA0020_0779	-1,63185329	ta20_049				IPR001623 DnaJ domain;IPR004640 Co-chaperone Hsc20;IPR009073 Co-chaperone HscB, C-terminal oligomerisation domain;IPR036386 HscB, C-terminal domain superfamily;IPR036869 Chaperone J-domain superfamily	O:(O) Posttranslational modification, protein turnover, chaperones
TA0020_0332	-1,6837151	ta20_011		3.4.24.59		IPR001567 Peptidase M3A/M3B catalytic domain;IPR024077 Neurolysin/Thimet oligopeptidase, domain 2;IPR024079 Metallopeptidase, catalytic domain superfamily;IPR045090 Peptidase M3A/M3B	O:(O) Posttranslational modification, protein turnover, chaperones
TA0020_1261	-1,77359329	ta20_038	CAZy	2.3.1.50	GT20	IPR004839 Aminotransferase, class I/classII;IPR015421 Pyridoxal phosphate-dependent transferase, major domain;IPR015422 Pyridoxal phosphate-dependent transferase, small domain;IPR015424 Pyridoxal phosphate-dependent transferase	O:(O) Posttranslational modification, protein turnover, chaperones
TA0020_0188	-1,89179261	ta20_030		2.5.1.18		IPR004045 Glutathione S-transferase, N-terminal;IPR036249 Thioredoxin-like superfamily;IPR036282 Glutathione S-transferase, C-terminal domain superfamily;IPR040079 Glutathione Transferase family	O:(O) Posttranslational modification, protein turnover, chaperones
TA0020_1099	-1,65614449	ta20_012	CAZy_TR		GT1	IPR000109 Proton-dependent oligopeptide transporter family;IPR018456 PTR2 family proton/oligopeptide symporter, conserved site;IPR036259 MFS transporter superfamily	P:(P) Inorganic ion transport and metabolism
TA0020_0277	-1,72010834	ta20_026		3.1.3.41		IPR006349 2-phosphoglycolate phosphatase, eukaryotic;IPR006357 HAD-superfamily hydrolase, subfamily IIA;IPR023214 HAD superfamily;IPR036412 HAD-like superfamily	P:(P) Inorganic ion transport and metabolism
TA0020_0368	3,836196638	ta20_056		2.1.1.229		IPR013216 Methyltransferase type 11;IPR029063 S-adenosyl-L-methionine-dependent methyltransferase	Q:(Q) Secondary metabolites biosynthesis, transport and catabolism

TA0020_1125	2,050633026	ta20_021	CAZy_TR		AA9	IPR003439 ABC transporter-like, ATP-binding domain;IPR003593 AAA+ ATPase domain;IPR011527 ABC transporter type 1, transmembrane domain;IPR017871 ABC transporter-like, conserved site;IPR027417 P-loop containing nucleoside triphosphate hydrolase;IPR036640 ABC transporter type 1, transmembrane domain superfamily	Q:(Q) Secondary metabolites biosynthesis, transport and catabolism
TA0020_1229	-1,88976267	ta20_050		1.14.14.154		IPR001128 Cytochrome P450;IPR002401 Cytochrome P450, E-class, group I;IPR036396 Cytochrome P450 superfamily	Q:(Q) Secondary metabolites biosynthesis, transport and catabolism
TA0020_0369	3,99973516	ta20_056	TF			IPR001005 SANT/Myb domain;IPR009057 Homeobox-like domain superfamily;IPR017930 Myb domain	S:(S) Function unknown
TA0020_1030	2,876209893	ta20_013				IPR019481 Transcription factor TFIIIC, triple barrel domain	S:(S) Function unknown
TA0020_0374	2,591722349	ta20_056	TF			IPR004875 DDE superfamily endonuclease domain;IPR006600 HTH CenpB-type DNA-binding domain;IPR007889 DNA binding HTH domain, Psq-type;IPR009057 Homeobox-like domain superfamily;IPR038903 Allergen Asp f 4	S:(S) Function unknown
TA0020_0373	2,583100546	ta20_056				IPR010733 Domain of unknown function DUF1308	S:(S) Function unknown
TA0020_1004	2,110465566	ta20_013	SM			IPR012535 Cell division protein Cdc14	S:(S) Function unknown
TA0020_0717	1,760601608	ta20_041				IPR010636 Cerato-ulmin hydrophobin family;IPR036686 Hydrophobin superfamily	S:(S) Function unknown
TA0020_0656	1,644224849	ta20_001				IPR029063 S-adenosyl-L-methionine-dependent methyltransferase	S:(S) Function unknown
TA0020_0681	1,505897625	ta20_053				IPR031349 General transcription factor TFIIH, subunit Tfb6	S:(S) Function unknown
TA0020_0891	-1,60776602	ta20_010	SM			IPR018820 Brefeldin A sensitivity protein-related, domain of unknown function DUF2421;IPR018823 Putative ER transporter, 6TM, N-terminal	S:(S) Function unknown
TA0020_1100	-1,64510991	ta20_012	TF			IPR001138 Zn(2)-C6 fungal-type DNA-binding domain;IPR021858 Fungal transcription factor;IPR036864 Zn(2)-C6 fungal-type DNA-binding domain superfamily	S:(S) Function unknown
TA0020_0593	-1,65857892	ta20_034	TF			IPR013087 Zinc finger C2H2-type;IPR036236 Zinc finger C2H2 superfamily	S:(S) Function unknown
TA0020_1040	-1,70539388	ta20_025				IPR020164 Cytochrome c oxidase assembly protein COX16	S:(S) Function unknown
TA0020_1154	-1,75934775	ta20_043				IPR007217 Per1-like	S:(S) Function unknown

TA0020_0483	-1,76415097	ta20_023			IPR010636 Cerato-ulmin hydrophobin family;IPR036686 Hydrophobin superfamily	S:(S) Function unknown
TA0020_0170	-1,82030257	ta20_004	CAZy_TR	GH28	IPR011701 Major facilitator superfamily;IPR020846 Major facilitator superfamily domain;IPR036259 MFS transporter superfamily	S:(S) Function unknown
TA0020_0023	-1,82358202	ta20_052	TF		IPR001138 Zn(2)-C6 fungal-type DNA-binding domain;IPR036864 Zn(2)-C6 fungal-type DNA-binding domain superfamily	S:(S) Function unknown
TA0020_0379	-1,93981061	ta20_056			IPR011990 Tetratricopeptide-like helical domain superfamily;IPR019734 Tetratricopeptide repeat;IPR044244 Tetratricopeptide repeat protein TTC27/Emw1	S:(S) Function unknown
TA0020_0284	-2,12891337	ta20_026			IPR000210 BTB/POZ domain;IPR011333 SKP1/BTB/POZ domain superfamily	S:(S) Function unknown
TA0020_0951	-2,24797878	ta20_019			IPR009959 Polyketide cyclase SnoaL-like;IPR032710 NTF2-like domain superfamily	S:(S) Function unknown
TA0020_0181	-2,31701008	ta20_030			IPR010829 Cerato-platanin;IPR036908 RlpA-like domain superfamily	S:(S) Function unknown
TA0020_0672	-2,35636948	ta20_053			IPR011989 Armadillo-like helical;IPR013180 Beta-catenin-like protein 1, N-terminal;IPR016024 Armadillo-type fold;IPR039678 Beta-catenin-like protein 1	S:(S) Function unknown
TA0020_0224	-2,50274763	ta20_029			IPR008030 NmrA-like domain;IPR036291 NAD(P)-binding domain superfamily;IPR045312 Phenylcoumaran benzylic ether reductase-like	S:(S) Function unknown
TA0020_0974	-2,5036072	ta20_019			IPR005624 Haem degrading protein HbpS-like;IPR010371 Uncharacterised protein family UPF0303;IPR038084 Haem degrading protein HbpS-like superfamily	S:(S) Function unknown
TA0020_0033	-2,58121933	ta20_052			IPR001810 F-box domain	S:(S) Function unknown
TA0020_0730	-2,6960304	ta20_020			IPR025702 Haem-containing dehydratase	S:(S) Function unknown
TA0020_1186	-3,0870818	ta20_024			IPR007834 DSS1/SEM1	S:(S) Function unknown
TA0020_0655	-3,3863725	ta20_001			IPR000250 Peptidase G1;IPR013320 Concanavalin A-like lectin/glucanase domain superfamily;IPR038656 Peptidase G1 superfamily	S:(S) Function unknown
TA0020_0489	-3,93769011	ta20_023	SM		IPR029058 Alpha/Beta hydrolase fold	S:(S) Function unknown

TA0020_1027	1,64759699	ta20_013				IPR001164 Arf GTPase activating protein;IPR009060 UBA-like superfamily;IPR015940 Ubiquitin-associated domain;IPR037278 ARFGAP/RecO-like zinc finger;IPR038508 ArfGAP domain superfamily	T:(T) Signal transduction mechanisms
TA0020_1039	-1,52413782	ta20_025				IPR000253 Forkhead-associated (FHA) domain;IPR008984 SMAD/FHA domain superfamily	T:(T) Signal transduction mechanisms
TA0020_0326	-2,05555114	ta20_011				IPR000648 Oxysterol-binding protein;IPR018494 Oxysterol-binding protein, conserved site;IPR037239 Oxysterol-binding protein superfamily	T:(T) Signal transduction mechanisms
TA0020_0419	-1,5659019	ta20_018				IPR000697 WH1/EVH1 domain;IPR003124 WH2 domain;IPR011993 PH-like domain superfamily;IPR033927 WASP family, EVH1 domain	T:(T) Signal transduction mechanisms;Z:(Z) Cytoskeleton
TA0020_0458	-1,72993937	ta20_028				IPR006895 Zinc finger, Sec23/Sec24-type;IPR006896 Sec23/Sec24, trunk domain;IPR006900 Sec23/Sec24, helical domain;IPR007123 Gelsolin-like domain;IPR012990 Sec23/Sec24 beta-sandwich;IPR029006 ADF-H/Gelsolin-like domain superfamily;IPR036174 Zinc finger, Sec23/Sec24-type superfamily;IPR036175 Sec23/Sec24 helical domain superfamily;IPR036180 Gelsolin-like domain superfamily;IPR036465 von Willebrand factor A-like domain superfamily	U:(U) Intracellular trafficking, secretion, and vesicular transport
TA0020_0673	-1,93620196	ta20_053		3.1.1.3		IPR005124 Vacuolar (H+)-ATPase G subunit	U:(U) Intracellular trafficking, secretion, and vesicular transport
TA0020_1121	1,501203645	ta20_021	CAZy		CBM18	IPR003593 AAA+ ATPase domain;IPR004273 Dynein heavy chain region D6 P-loop domain;IPR013594 Dynein heavy chain, tail;IPR013602 Dynein heavy chain, linker;IPR024317 Dynein heavy chain, AAA module D4;IPR024743 Dynein heavy chain, coiled coil stalk;IPR026983 Dynein heavy chain;IPR027417 P-loop containing nucleoside triphosphate hydrolase;IPR035699 Dynein heavy chain, hydrolytic ATP-binding dynein motor region;IPR035706 Dynein heavy chain, ATP-binding dynein motor region;IPR041466 Dynein heavy chain, AAA 5 extension domain;IPR041658 Dynein heavy chain AAA lid domain;IPR042219 Dynein heavy chain AAA lid domain superfamily;IPR042222 Dynein heavy chain, domain 2, N-terminal;IPR042228 Dynein heavy chain, linker, subdomain 3;IPR043157 Dynein heavy chain, AAA1 domain, small subdomain	Z:(Z) Cytoskeleton

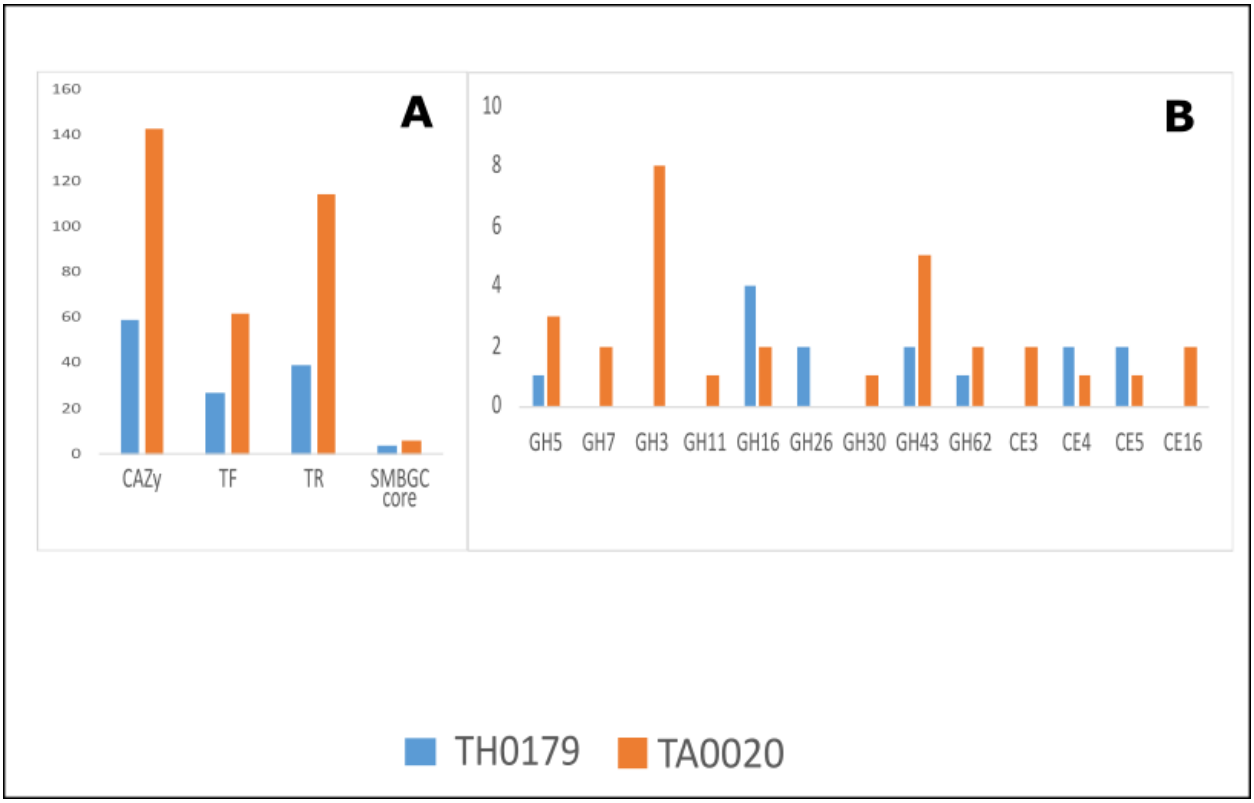
818

Table 4. DNA statistics for alignment against Trichoderma references

Strain	TrQM6a	ThCBS226.95	TaIMI 206040
TA0020	7,46%	10.42%	87.34%
TH0179	12,31%	95,47%	9,71%

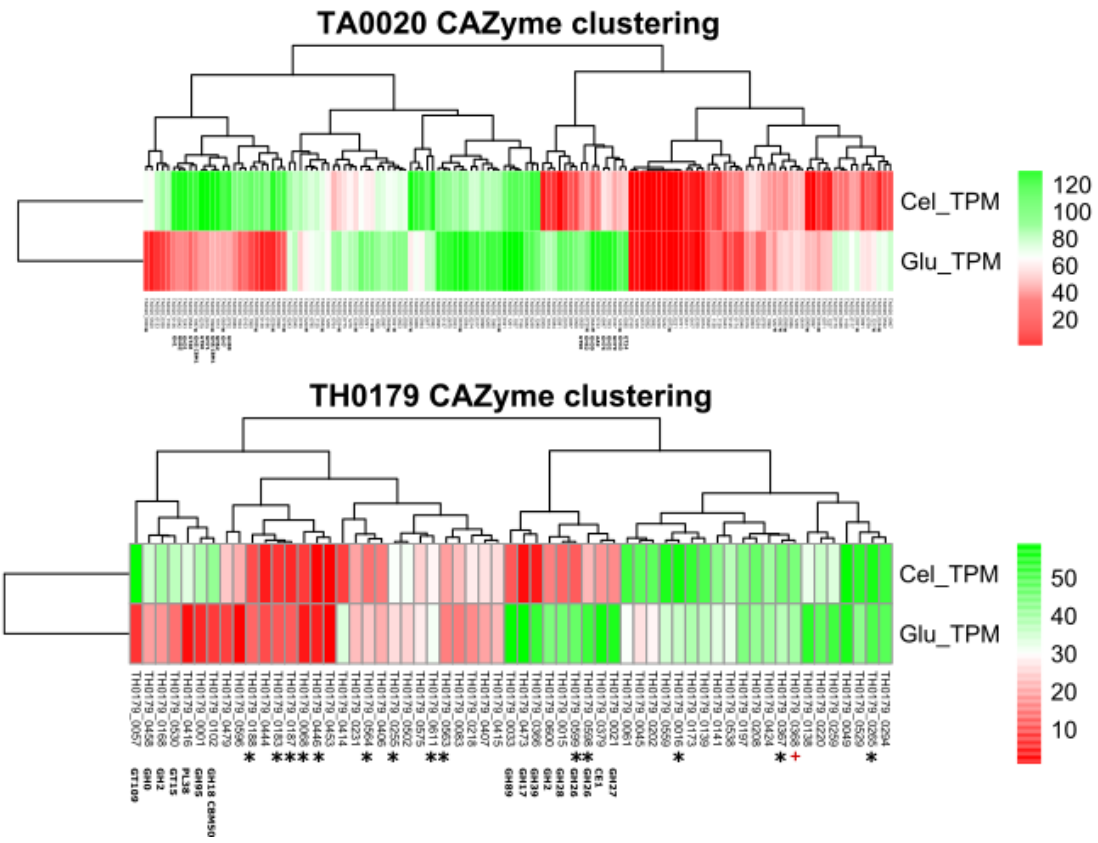
Figures

Fig. 1



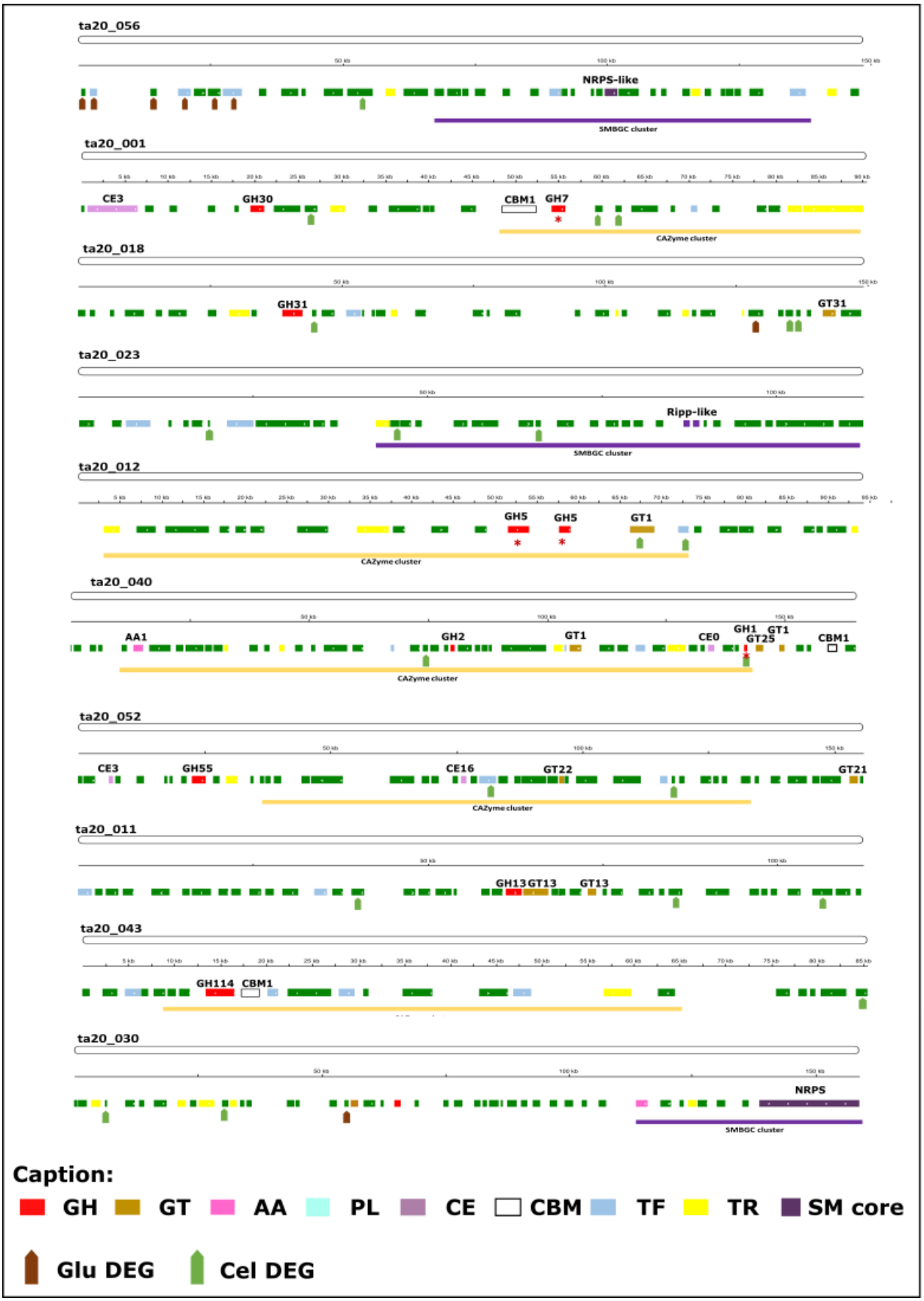
Caption: A direct comparison of relevant categories between TA0020 and TH0179**. B** Comparison of lignocellulolytic CAZyme families between TA0020 and TH0179.

Fig. 2



Caption: TA0020 and TH0179 CAZyme clustering analysis comparing cellulose and glucose degradative conditions. * Represents lignocellulolytic CAZyme families from Andlar et al. 2022, and the red + identifies the expression outlier from TH0179.

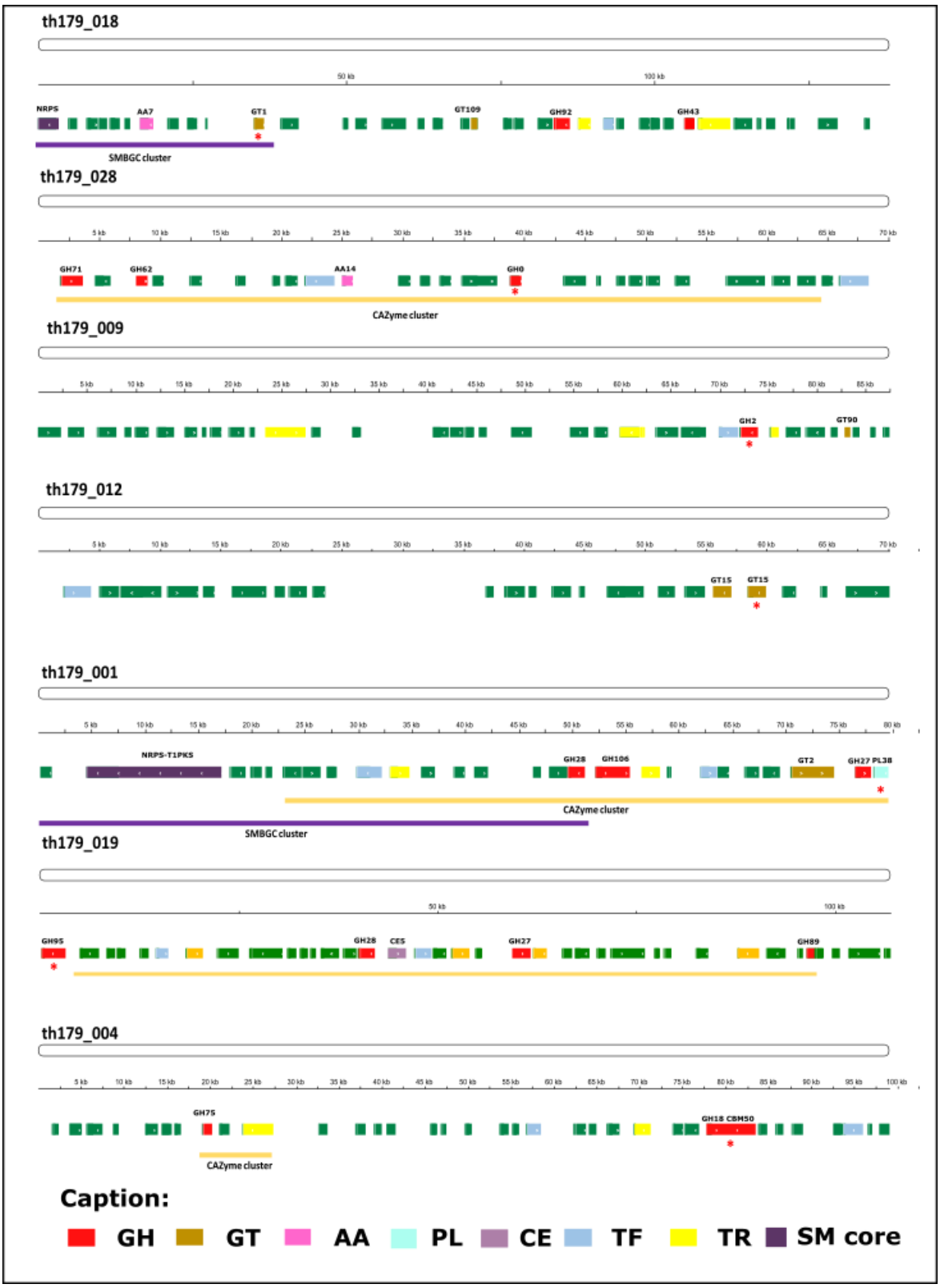
Fig 3.


ta20_056
NRPS-like
SMBGC cluster
ta20_001
CE3
GH30
CBM1
GH7
CAZyme cluster
ta20_018
GH31
GT31
ta20_023
Ripp-like
SMBGC cluster
ta20_012
GH5
GH5
GT1
CAZyme cluster
ta20_040
AA1
GH2
GT1
CE0
GH1
GT1
GT25
CBM1
CAZyme cluster
ta20_052
CE3
GH55
CE16
GT22
GT21
CAZyme cluster
ta20_011
GH13
GT13
GT13
ta20_043
GH114
CBM1
ta20_030
NRPS
SMBGC cluster
Caption:
GH
GT
AA
PL
CE
CBM
TF
TR
SM core
Glu DEG
Cel DEG

832

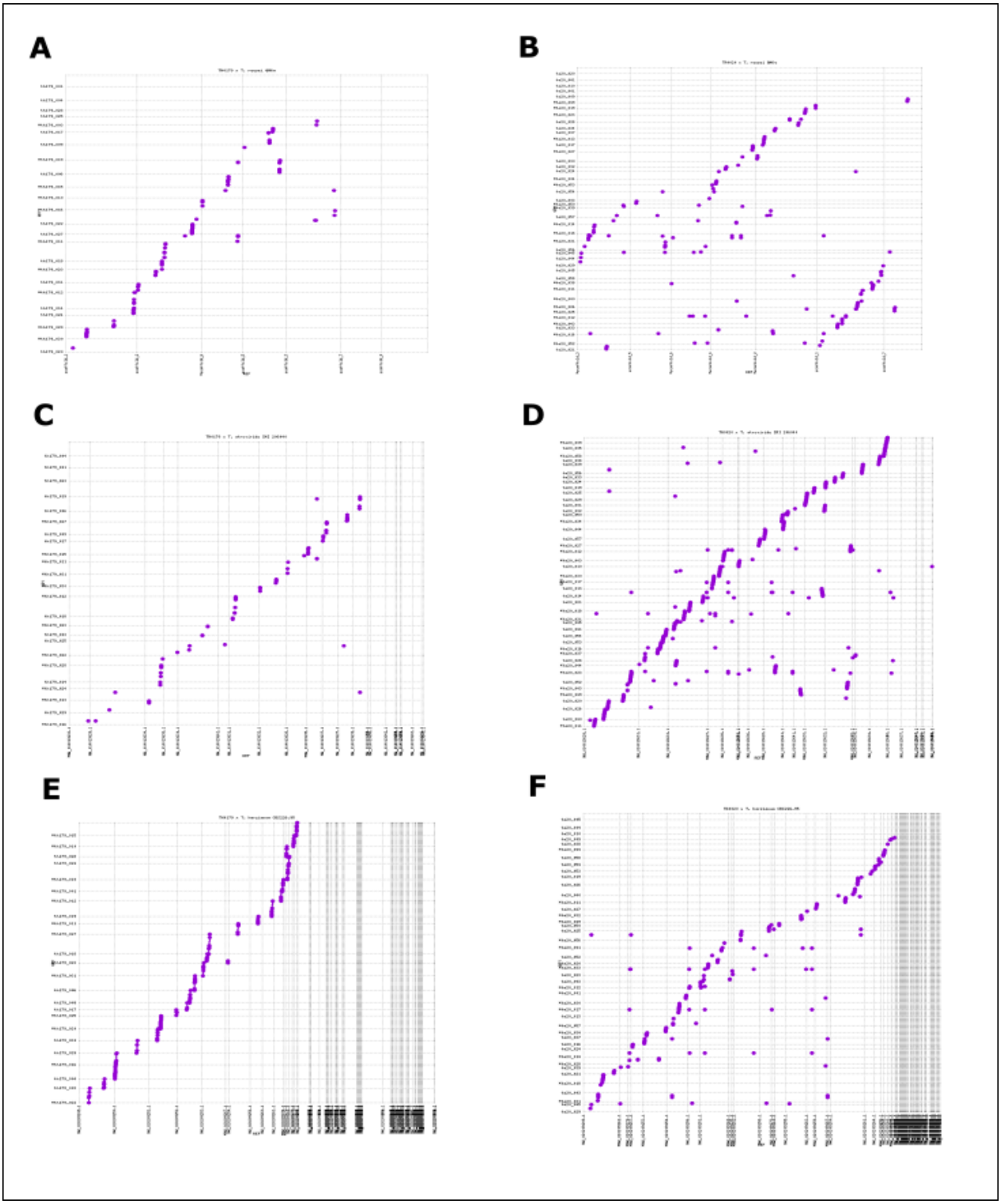
Caption: TA0020 IGV visualization of major degradative regions identified via CAZyme expression clustering analysis and differential expression analysis. The red * indicates the CAZyme identified by clustering analysis.

Fig. 4


th179_018
NRPS
AA7
GT1
GT109
GH92
GH43
SMBGC cluster
th179_028
GH71
GH62
AA14
GH0
CAZyme cluster
th179_009
GH2
GT90
th179_012
GT15
GT15
th179_001
NRPS-T1PKS
GH28
GH106
GT2
GH27 PL38
CAZyme cluster
SMBGC cluster
th179_019
GH95
GH28
CE5
GH27
GH89
th179_004
GH75
GH18 CBM50
CAZyme cluster
Caption:
GH
GT
AA
PL
CE
TF
TR
SM core

Caption: TH0179 IGV visualization of major cellulolytic genomic regions identified via CAZyme expression clustering analysis. The red * indicates the CAZyme identified by clustering analysis. The red + indicates the outlier CAZyme.

Fig. 5



Caption: Alignments visualized using a MUMmer dot plot with the parameters Minmatch 50 bp and match length 500 bp against *T. reesei* QM6a, *T. harzianum* CBS22695 and *T. atroviride* IMI 206040. **A** TH0179 x *T. reesei* QM6a. **B** TA0020 x *T. reesei* QM6a**. C** TH0179 **x** *T. atroviride* IMI206040. **D** TA0020 **x** *T. atroviride* IMI206040**. E** TH0179 **x** *T. harzianum* CBS226.95 **F** TA0020 **x** *T. harzianum* CBS226.95.

850